MEGASAT: automated inference of microsatellite genotypes from sequence data

Article (Accepted Version)

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) doi: 10.1111/1755-0998.12561

MEGASAT: automated inference of microsatellite genotypes from sequence data

LUYAO ZHAN,*[1] IAN G. PATERSON,†[1] BONNIE A. FRASER,‡ BETH WATSON,† IAN R. BRADBURY,§ PRAVEEN NADUKKALAM RAVINDRAN,* DAVID REZNICK,¶ ROBERT G. BEIKO* and PAUL BENTZEN†

*Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax Nova Scotia B3H 4R2, Canada, †Marine Gene Probe Laboratory, Department of Biology, Dalhousie University, 1355 Oxford Street, Halifax Nova Scotia B3H 4R2, Canada, ‡Evolution Behaviour and Environment Group, University of Sussex, Sussex House, Falmer, Brighton BN1 9RH, UK, §Salmonids Section, Science Branch, Department of Fisheries and Oceans Canada, 80 East White Hills Road, St. John's Newfoundland A1C 5X1, Canada, ¶Department of Biology, University of California, Riverside, CA 92521, USA

Abstract

MEGASAT is software that enables genotyping of microsatellite loci using next-generation sequencing data. Microsatellites are amplified in large multiplexes, and then sequenced in pooled amplicons. MEGASAT reads sequence files and automatically scores microsatellite genotypes. It uses fuzzy matches to allow for sequencing errors and applies decision rules to account for amplification artefacts, including nontarget amplification products, replication slippage during PCR (amplification stutter) and differential amplification of alleles. An important feature of MEGASAT is the generation of histograms of the length–frequency distributions of amplification products for each locus and each individual. These histograms, analogous to electropherograms traditionally used to score microsatellite genotypes, enable rapid evaluation and editing of automatically scored genotypes. MEGASAT is written in Perl, runs on Windows, Mac OS X and Linux systems, and includes a simple graphical user interface. We demonstrate MEGASAT using data from guppy, *Poecilia reticulata*. We genotype 1024 guppies at 43 microsatellites per run on an Illumina MiSeq sequencer. We evaluated the accuracy of automatically called genotypes using two methods, based on pedigree and repeat genotyping data, and obtained estimates of mean genotyping error rates of 0.021 and 0.012. In both estimates, three loci accounted for a disproportionate fraction of genotyping errors; conversely, 26 loci were scored with 0–1 detected error (error rate ≤0.007). Our results show that with appropriate selection of loci, automated genotyping of microsatellite loci can be achieved with very high throughput, low genotyping error and very low genotyping costs.

Keywords: animal mating/breeding systems, bioinformatics/phyloinformatics, captive populations, conservation genetics, landscape genetics, population genetics – empirical

Received 18 April 2016; revision received 31 May 2016; accepted 31 May 2016

Introduction

Microsatellites, by virtue of their abundance in all eukaryotic organisms, high levels of polymorphism and relatively easy assay have been the most widely applied molecular genetic markers in molecular ecology and many other fields of biology over the last two decades (e.g. Wright & Bentzen 1994; Jarne & Lagoda 1996; Provan *et al.* 2001; Putman & Carbone 2014). In recent years, single nucleotide polymorphisms (SNPs) have gained in popularity relative to microsatellites (Guichoux *et al.* 2011; Putman & Carbone 2014). This trend reflects some widely cited advantages of SNPs over microsatellites, such as greater abundance in genomes, lower genotyping error rates, greater amenability to high-throughput genotyping and, potentially, lower cost per single-locus genotype (e.g. Guichoux *et al.* 2011). However, these advantages of SNPs are not always realized or relevant when fewer than thousands of loci are required. For a variety of applications including analyses of linkage disequilibrium, association, parentage, kinship, individual identity, population expansions and contractions (bottlenecks) and genetic structure, multi-allelic microsatellites, on a per-locus basis, are 2–20× more

Correspondence: Paul Bentzen, Fax: +902-494-3736; E-mail: paul.bentzen@dal.ca

[1]These authors contributed equally to this work.

powerful than SNPs (Haasl & Payseur 2010; Guichoux *et al.* 2011; Aime *et al.* 2014). For sibship reconstruction, the relative advantage of microsatellites is essentially infinite, because such analyses require a minimum of four alleles at informative loci (e.g. Jones & Wang 2010). In addition, the mutational properties of microsatellites make them much less prone to ascertainment bias and give them unique potential as reliable, 'fast' molecular clocks (Li *et al.* 2008; Sun *et al.* 2009).

The efficiencies and economies associated with SNP genotyping are best realized at large scales: many loci (minimally, hundreds) genotyped in many individuals. Such large-scale genotyping efforts require large initial investments in set-up costs, which may not be cost-effective when experimental needs require only more modest numbers of loci (but see Campbell *et al.* 2014). Small-scale genotyping of SNPs can be more costly than genotyping of microsatellites, particularly when the lower information content per locus is considered. Although the use of technologies such as microfluidic devices can lower the cost of small-scale SNP assays, these require access to expensive and specialized instrumentation (Seeb *et al.* 2009).

The most important disadvantage associated with microsatellite genotyping stems from the traditional reliance on electrophoretic methods and the necessarily imperfect inference of genotypes from DNA fragment mobility data. In spite of the development of multiplex and semi-automated microsatellite genotyping using electrophoresis-based DNA analyzers (Kimpton *et al.* 1993), reliance on electrophoresis remains a limiting factor in microsatellite genotyping. Challenges in data interpretation include distinguishing alleles from amplification artefacts, resolving alleles that can differ by as little as a single base pair in size, allele size ambiguities caused by 3′ adenylation of PCR products and detecting weakly amplifying alleles. Standardizing allele size calls among different individuals, laboratories and electrophoretic platforms is also a significant challenge (Moran *et al.* 2006; Guichoux *et al.* 2011). A notable example of the last point is that the inferred sizes of microsatellite alleles genotyped on slab gels and capillary systems often differ (Moran *et al.* 2006).

Next-generation DNA sequencing (NGS) offers a powerful alternative to electrophoresis for the analysis of microsatellite genotypes. Sequencing read lengths for some NGS systems (e.g. Illumina MiSeq: 300b and Thermo Fisher Ion Torrent: 400b) encompass the range of allele sizes of most microsatellite loci currently genotyped with electrophoresis. This indicates the potential to directly read microsatellite genotypes from amplicon sequence data. Potential benefits include much greater throughput, lower consumable costs, and greater accuracy in genotyping, as inferring genotypes directly from sequence data avoids all of the artefacts associated with electrophoretic detection. To realize the potential of these methods, suitable software is needed to convert raw amplicon sequence data to multilocus microsatellite genotypes. Such software needs to deal with a variety of artefacts that can occur during PCR amplification and sequencing of microsatellites. Chief among these is amplification stutter, in which replication slippage during PCR produces additional amplification products that differ from the 'true' allele length by multiples of the microsatellite repeat unit. Additional artefacts to address include differential amplification of alleles and allelic 'dropout' caused by amplification bias favouring small alleles, or low DNA template quantity or quality. The software also needs to cope with sequencing errors that could interfere with identification of microsatellite loci within pooled amplicon libraries.

Recently, Suez *et al.* (2016) described a method for genotyping microsatellites from NGS data. Their method builds a theoretical parametric model for genotypes and aims to find the optimized parameters via minimizing the squared difference between the observed length distribution and theoretical parametric model. Their method considers only the repeat array portion of the microsatellite and only those consisting of pure (i.e. not compound or interrupted) arrays. However, this method suffers from some disadvantages. If the parametric model cannot correctly simulate the mode of data, it could induce bias into the inferred results. Furthermore, parametric modelling always comes with a high computational cost. We developed a method that uses a very different approach. Our method includes flanking sequences in the genotype, does not require pure repeat arrays and employs a much less computationally intensive approach that uses sequence depth ratios and decision rules to infer genotypes. Here, we present MEGASAT, new software that allows the rapid conversion of DNA sequence data from highly multiplexed and pooled microsatellite amplicons to multilocus genotypes. MEGASAT has three primary functions: (i) demultiplex highly multiplexed NGS data (FASTQ or FASTA) into locus-specific files, based on primer and flanking sequences; (ii) automate the scoring of microsatellite genotypes, using sequence depth with decision rules to account for amplification artefacts; (iii) generate plot files (histograms of sequence length-frequency distributions) for manual verification of genotypes. MEGASAT outputs predicted multilocus genotypes to tab-delimited text files that can be imported into spreadsheets. The plot file histograms are analogues of the electropherograms traditionally used to interpret microsatellite genotypes obtained with capillary electrophoresis data, enable rapid data checking and editing of automated genotype calls. MEGASAT is implemented in the Perl language and can be used either from

a command line or via a graphical user interface (GUI) in Windows and Mac OS X.

We demonstrate the application of MEGASAT for microsatellite genotyping using multiplexed, pooled amplicons of 43 guppy (*Poecilia reticulata*) microsatellites sequenced using Illumina MiSeq. We further demonstrate a high level of reproducibility and accuracy of MEGASAT-called microsatellite genotypes by a combination of repeated genotyping of independently extracted and amplified duplicate samples, and examination of known pedigrees of guppies to identify genotyping errors.

Methods

Laboratory

MSATCOMMANDER (Faircloth 2008) was used to select di- and trinucleotide microsatellites from the guppy genome (NCBI BioProject PRJNA238429) that met the following criteria: >6 repeat units and predicted amplicon size 60–158 bp. Of 2915 loci that met these criteria, 448 loci with >7 repeats were chosen for further analysis (Appendix S1, Supporting information). Oligonucleotides were purchased from Integrated DNA technologies (IDT, Coralville, IA, USA). Forward and reverse microsatellite primers were tailed with Illumina (San Diego, CA, USA) Read1_(CCCTACACGACGCTCTTCC GATCT) and Read2_(GTTCAGACGTGTGCTCTTCCG ATCT) sequencing primers, respectively, resulting in oligonucleotides 42–47b length.

In initial trials, loci were amplified in 10-locus multiplex PCRs. Subsequent libraries were created using one 43-locus multiplex per sample (see Appendix. S3 for summary stats, Supporting information). Multiplex PCRs were performed in 3.5 μL volumes using Qiagen (Venlo, the Netherlands) Type-IT 2× Mastermix (1.75 μL), 0.2 μM each oligonucleotide and 0.7 μL genomic DNA (estimated to be ~275 pg). PCRs were conducted on Eppendorf (Hamburg, Germany) Mastercycler ep384 PCR machines using the following parameters: 94 °C for 15 min, followed by 20 cycles of 94 °C 30 s, 57 °C, 180 s, 72 °C 60 s, with a final extension at 68 °C for 30 min.

Indexing sequences were added to the PCR products using a second PCR. The index PCR used oligonucleotides composed of Illumina annealing adapter sequences, a 6b index (barcode) and the Illumina sequencing primers. We used 32 Index_1 oligonucleotides and 32 Index_2 oligonucleotides to differentiate 1024 individuals in each MiSeq sequencing run. Indexing PCRs were performed in 5 μL total volume with 0.25 U *Taq* DNA polymerase (New England Biolabs, Ipswich, MA, USA), 0.5 μL Thermopol 10× buffer (NEB), 0.2 mM each dNTP, 0.2 μM Index_1 oligo, 0.2 μM Index_2 oligo and 0.3 μL of 20-fold diluted multiplex-PCR product. Cycling parameters were as follows: 95 °C 2 min, followed by 18 cycles of 95 °C 20 s, 60 °C 60 s, 72 °C 60 s with a final extension at 72 °C for 10 min.

Indexed PCR products were pooled and cleaned using Ampure XP (Beckman Coulter, Pasadena CA, USA) or Sera-Mag Speedbeads (GE Healthcare, Little Chalfont, UK) magnetic beads (1.8:1 bead:DNA library ratio). Libraries were quantified using Kapa (Wilmington, MA, USA) Library Quantification for Illumina on a Roche (Basel, Switzerland) LC480 qPCR instrument following manufacturers' protocols. Libraries were sequenced at 10–12 pM concentration using MiSeq v3 chemistry with 150 cycles in one direction and dual indexing. Indexed individuals were demultiplexed with the MISEQ SEQUENCE ANALYSIS software. Prior to developing MEGASAT, we used the GENEIOUS R7 software (Kearse *et al.* 2012) 'separate by barcode' function to demultiplex loci within an individual, and then microsatellite genotypes were scored using the depth histograms generated within GENEIOUS. Once MEGASAT was working, we used GENEIOUS to verify the performance of MEGASAT.

As more sequencing runs were performed and we learned more about each locus, we refined the process by dropping loci that had low information content, evidence of nulls or inability to multiplex well. For long-term data collection, we settled on 43 loci that we multiplex in a single PCR, using the same reaction conditions as our initial PCRs. Initial experiments used the Illumina v2 chemistry (300 cycle kits); however, the majority of libraries were sequenced using the v3 chemistry (150 cycles). Theoretical depths per locus per individual based on the minimum MiSeq performance specifications for v3 single read chemistry are 22 M reads/(1024 individuals * 43 loci) = 500 reads. Our actual average depth per locus per individual was 388 (std154). Comments on genotyping performance in typical sequencing runs are presented in Appendix S2 (Supporting information).

Software

Microsatellite-containing amplicon sequences have the following components: forward primer (FP), forward flank (FF), microsatellite repeat array (MRA), reverse flank (RF) and reverse primer complement (RP) (Fig. 1). MEGASAT uses reference data for each microsatellite locus to identify sequences associated with individual loci and to remove primer sequences. The FF and RF portions of the microsatellite amplicon are retained as part of the allele, for two reasons: (i) The flank sequences may contain insertions or deletions (indels) that contribute to allelic diversity and (ii) The boundaries of the MRA may not be clear in some loci; retaining the two flanking

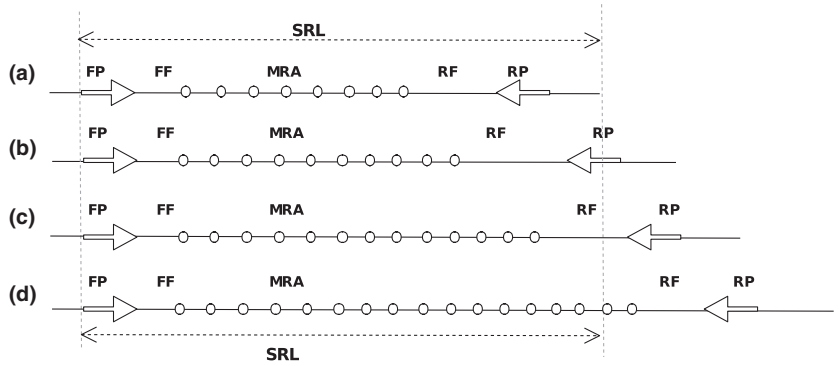

Fig. 1 Schematic showing a microsatellite amplicon. In (a), the forward primer (FP), forward flank (FF), microsatellite repeat array (MRA), reverse flank (RF) and reverse primer reverse-complement (RP) are present within the sequence read length (SRL). In (b), a longer MRA leaves only a few bases of the 3′ end of the RP within the SRL. In (c), an even longer MRA causes only part of the RF to be present with the SRL. In (d), the MRA extends past the end of the SRL. In cases (a) and (b), MEGASAT is able to detect the 3′ end of the RP and directly ascertain the length of the amplified microsatellite allele, which consists of FF + MRA + RF. In case (c), MEGASAT detects the end of the MRA and adds the reference length for the RF to infer the allele length. In case (d), MEGASAT detects the length of the FF and adds an integer value, to denote alleles that exceed the length of the SRL. See text for further details.

sequences avoids the need in most cases to define exact boundaries for the MRA, although our script includes the ability to define the boundary of the MRA when needed (see below).

The first function of MEGASAT is to sort the input reads (all sequences for a given sample) into per-locus files containing only those reads of interest by discarding those which do not contain the locus-specific priming and flanking sequence. The process of identifying and trimming off primers may be complicated by one or more factors, including sequencing errors and the possibility that all or part of the reverse primer complement, or even all or part of the RF, may be absent from the sequenced portion of the amplicon. This can occur when the size of the amplified microsatellite allele exceeds the read length of the sequencing chemistry being used (Fig. 1b–d).

Sequencing errors may cause a microsatellite-containing sequence to be erroneously discarded because of one or more mismatches between the reference primer (or microsatellite flank) sequence and the reference sequence used to identify the locus. To overcome this problem, MEGASAT allows a tolerance for mismatches when matching reference sequences for primers and flanking regions with observed sequences. The number of allowed mismatches is a user-controlled variable. The function finds the starting position of a near-exact match in the target sequence, which can be used to enable trimming of primers.

Another important function in MEGASAT helps to find the end of the MRA. The function incorporates tolerance for sequencing errors or SNPs (i.e. 'fuzzy' matching) in one or two microsatellite repeat units. This function can also be used to find the end of the RF when only a few bases of the reverse primer complement (RP) are present in the sequence. MEGASAT uses the Hamming distance (the number of differences between two strings of equal length) to find the starting point of any reverse primer complement in a sequence when the complete reverse primer complement is not present in the sequence. We use this function because it allows the primers to be trimmed off at the correct position even if there are length variations in the RF. Figure 2 shows a detailed overview of the procedure MEGASAT used to evaluate candidate microsatellite-containing sequences from the input FASTQ or FASTA file. To be included in the set of trimmed sequences, an input sequence must contain a match to the FP and the FF. As long MRAs may be equal to or exceed the read length, MEGASAT accepts sequences if they contain complete or partial RP or RF sequences, or if the identified MRA extends to the end of the sequence read. In such cases, a large constant score is added to the read length to identify these alleles as a separate class that could not be distinguished due to their length. Primers are removed from retained sequences, while sequences that do not satisfy the above criteria are saved to a second file for inspection by the user.

The second major application of MEGASAT is to predict microsatellite genotypes. Once the genotyping set is complete, MEGASAT determines the lengths of all the trimmed sequences for each locus and each individual and records the count of sequence length variants for each locus in each individual. The next step is to infer allele sizes, and then genotypes, based on the length distribution of sequences obtained in the previous step. The process of inferring alleles and genotypes is complicated

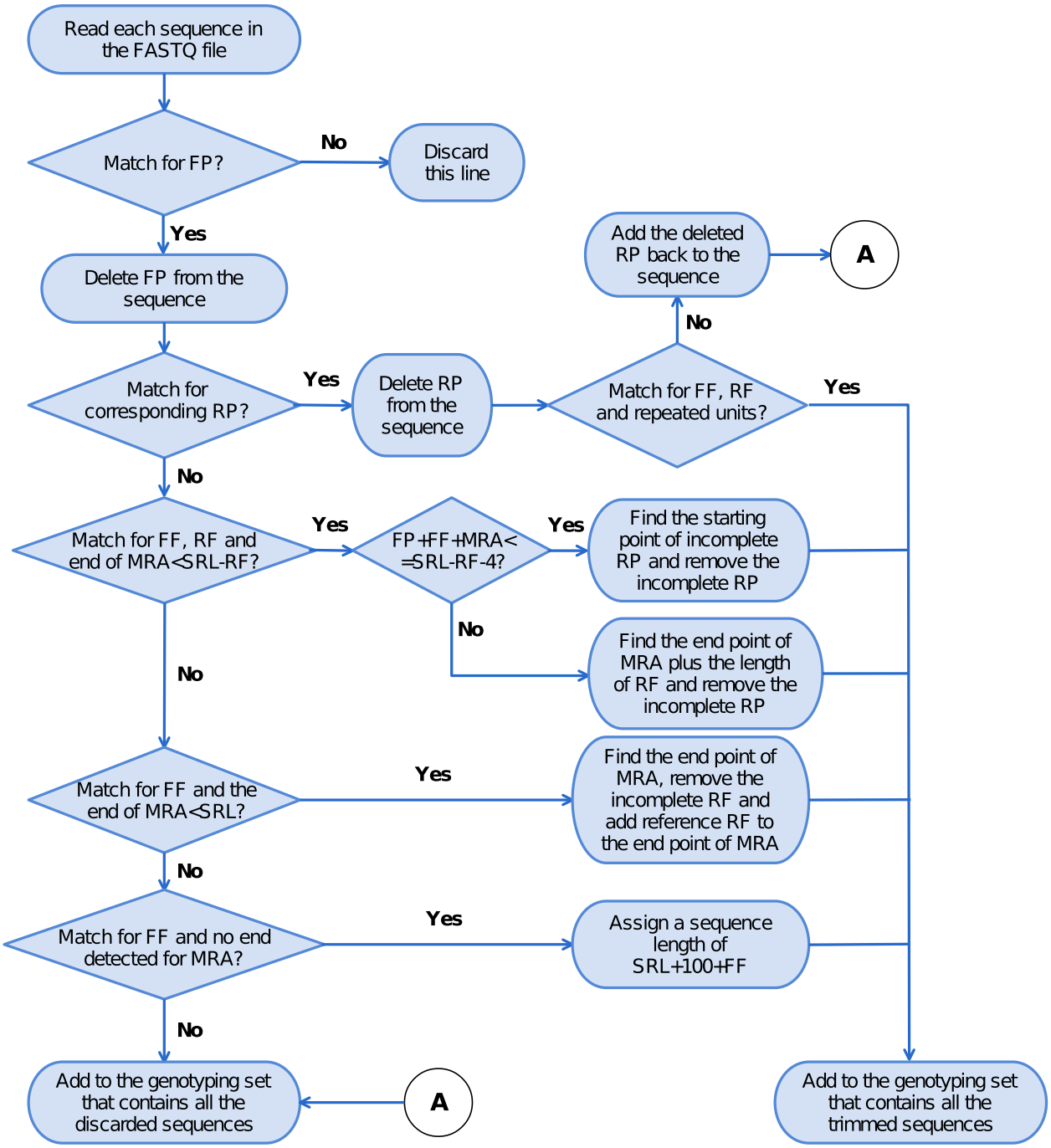

Fig. 2 Flow chart of the algorithm that MEGASAT uses to trim off microsatellite primers. The abbreviations for microsatellite amplicon components are the same as those given in Fig. 1.

by length artefacts that can arise during PCR amplification and sequencing. These include the following:

1 *Amplification 'stutter'*—This is the most common problem hindering the interpretation of microsatellite genotypes and occurs as replication slippage during PCR, resulting in products that are usually one to several repeat units shorter than the true allele size, but can occasionally extend upward in size from the correct allele length (such 'up-stutter' is more common with large alleles containing more repeat units).
2 *Large allele dropout*—The smaller of two alleles in heterozygotes is commonly amplified more strongly,

and the extent of the amplification bias usually increases with the difference in allele sizes.

3 *Stochastic allele dropout*—Occasionally, in heterozygotes, one allele is preferentially amplified over the other, but the amplification bias is not closely related to allele size. This can occur when amplification is from a small amount of template DNA, or the template DNA is degraded.

4 *Sequencing indels*—Spurious indels can arise as sequencing errors.

Amplification stutter and sequencing indels could result in artefacts being misinterpreted as microsatellite alleles, and all three amplification artefacts can cause heterozygotes to be mis-scored as homozygotes. MEGASAT employs a number of rules to distinguish true alleles from artefacts, and infer genotypes. The process of allele and genotype inference is shown in Fig. 3 and described briefly below. The process is outlined in more detail in Appendix S2 (Supporting information).

MEGASAT first ensures a minimum depth threshold, below which no allele calls are made. If the sum of the two most common sequence length variants exceeds the minimum threshold (default = 50), MEGASAT will score the genotypes. MEGASAT will correctly score most microsatellite genotypes, although we recommend review of the allele calls, especially in the early stages of a project when one is still characterizing the alleles for each locus. As expected, MEGASAT is most accurate when scoring typical, 'clean' microsatellites, that is samples with only one or two high-depth amplicons and a normal stutter pattern. MEGASAT is also capable of correctly scoring alleles for most atypical microsatellite amplification patterns. Note that 'atypical' does not mean

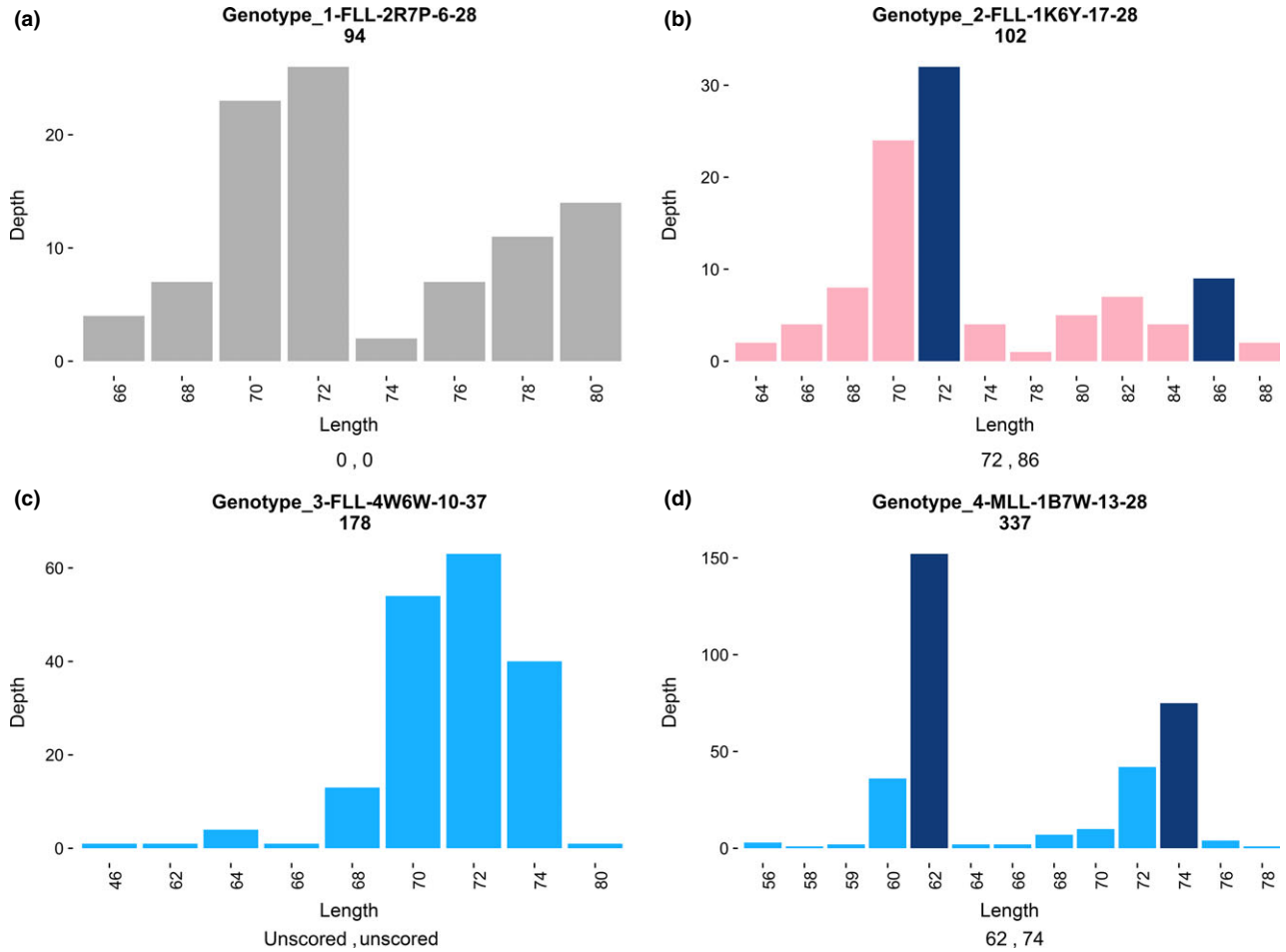

Fig. 3 Examples of MEGASAT portable document format (pdf) histograms that show the frequencies of sequence length variants per microsatellite locus per individual. Sample IDs title each plot, followed by the total depth. Genotypes are listed under the *x*-axis. Colour codes include (a) grey for samples below the minimum depth threshold, no alleles called. (b) pink warning that the depth is close to the minimum, deep blue indicates allele calls (72/86). (c) blue for acceptably high depths, no alleles called. (d) Deep blue indicates allele calls (62/74).

uncommon; in fact, in our experience, atypical patterns occur remarkably often.

The decision process regarding whether to accept an amplification product as a true allele is based on the depth ratios of as many as four of the most common length variants among amplification products relative to the most common length variant (which we term A1: see Appendix. S3 for example, Supporting information). The decision process considers the relative size (smaller or larger than A1) and the difference in size of potential alleles relative to the most abundant length variant. The important decision variables are user-definable (see User Manual for complete description) on a per-locus basis. The majority of loci score well with the default values, but some minor adjustments in the variable thresholds can improve scoring of specific loci.

The third major function of MEGASAT is to enable review of MEGASAT's genotype calls by (i) creating output files of genotyped sequences and discarded sequences and (ii) creating depth vs. size histogram plots for all samples per locus (Fig. 3). The genotyped and discarded sequences can be reviewed to ensure the decision variables are set correctly, and MEGASAT is not overly rigorous in accepting artefact sequences or discarding true alleles. These files are important early in a project when one is still characterizing the loci. The plots are a graphical representation of the allele calls MEGASAT has made and are an important tool for quickly reviewing the veracity of the genotypes. The plots are presented in pdf format. The plots are colour-coded for easy review. Grey indicates a sample below the minimum depth threshold and scored with a '0 0' genotype in the GENOTYPE.TXT file. Pink histograms serve as a visual clue that the depth is just marginally above the minimum threshold (<threshold +10) and blue indicates a high depth (>threshold +10). Allele calls are plotted in deep blue, allowing the reviewer to scan quickly over the plot files to see whether MEGASAT has called the alleles correctly. When an allele mis-call is identified, the reviewer can correct the call in the data file.

Estimation of genotyping error rates

Two approaches were used to evaluate the reliability of microsatellite genotypes obtained with MEGASAT. In one approach, 37 individuals were randomly resampled for tissue (scales). DNA extractions, PCR, sequencing and genotyping using MEGASAT were carried out independently, except that some repeat-genotyped individuals were sequenced in the same sequencing run. In the second approach, 71 guppies from known, laboratory-reared crosses were genotyped, and parent–offspring triads were examined for genotypes that would violate Mendelian rules of inheritance. In both approaches, genotyping error was evaluated for MEGASAT-scored genotypes both with and without additional manual editing of genotypes.

Results and discussion

MEGASAT processing time is determined by the size and complexity of the data set. Our NGS data sets from an Illumina MiSeq are typically comprised of 1024 FASTQ files, one per individual, with each file containing sequences for 43 microsatellite loci. Using ILLUMINA v3 chemistry, a data set of 8.6 GB (~28.5 million 150b reads) takes approximately 25.4 h to genotype on a Windows 7 PC with 8 GB RAM, or about 2.08 s/genotype. Creation of plot files following genotyping takes another ~1.5 h to create 43 PDFs (one per locus) containing 1024 samples each (44 032 plots).

Rates of genotyping error for MEGASAT-scored genotypes differed between the two methods used, testing for violation of Mendelian inheritance or repeat genotyping, and among loci (Appendix S4, Table S4, Supporting information). For MEGASAT-scored genotypes, the mean estimated error rate per allele was 0.021 for the pedigree-based method. Most of the genotyping errors detected using this method occurred with a few loci: three loci had error rates exceeding 0.1 (0.109–0.129). By contrast, 16 loci had no detected errors (error rate <0.007) and 10 loci had a single error (error rate ≈0.007); the remaining loci had intermediate error rates (Table S4, Supporting information).

Estimates of genotyping error obtained with the repeat genotyping method were lower (mean genotyping error = 0.012). The three most error-prone loci had estimated genotyping error rates of 0.040–0.050. Among the other loci, 18 loci had no detected genotyping errors and eight loci had a single error (error rate ≈ 0.007).

Using the histograms of sequence length variation produced for each single-locus genotype, we performed manual curation, which resulted in reduced mean genotyping error rates, particularly for those few loci that had the highest error rates in the automated genotype calls. Manual editing reduced the mean error rate from 0.021 to 0.010 in the pedigree-based estimates, and from 0.012 to 0.007 in the repeat genotyping-based estimates. In the pedigree-based estimates, manual editing substantially reduced genotyping error at the eight most error-prone loci; mean genotyping error rates for these eight loci were 0.087 and 0.023 before and after manual editing, respectively (Table S4, Supporting information). By contrast, manual editing produced no gains in accuracy for 28 loci, either because no errors were detectable in the automatically scored loci, or because the rare errors that did occur were scored the same way by a person and by MEGASAT. Results were similar with repeat genotyping-based error estimates, except that overall error rates, and the gains realized from manual editing, were smaller.

Manual editing reduced the mean error rate from 0.044 to 0.018 for the three most error-prone loci in this analysis, but produced little or no gain in accuracy for 35 loci, for the same reasons as before.

These results suggest some important considerations for microsatellite genotyping using MEGASAT. First, completely automated genotype prediction is feasible for many loci. In our experiments, automated genotype prediction resulted in mean error rates of 0.003–0.004 for 28 loci (estimated using either method), and no genotyping errors detected for 16–18 loci. Slightly higher mean rates of genotyping error occur with fully automated genotyping of up to 40 loci in our panel. Moreover, our panel of 43 loci were selected for their easy amplification in a large (43-plex) multiplex panel, high polymorphism and suitable allele size ranges, but they were not rigorously screened for their tendency to produce easily interpretable genotypes. A clear implication is that further screening of candidate microsatellite loci could have produced more loci that met all desired criteria, including amplification products amenable to highly accurate, fully automated scoring. We have recently adapted three 'legacy' sets of di- and tetranucleotide repeat microsatellite loci for other fish species (previously scored using electrophoresis) to our NGS-MEGASAT-based approach, with excellent results (data not shown). This experience, and the fact that Illumina MiSeq and Thermo Fisher Ion Torrent reads currently reach 300b and 400b lengths, respectively, suggests that a high proportion of the vast number of microsatellite markers that have already been developed could be adapted to this genotyping method.

Second, such fully automated genotype prediction brings great advantages in genotyping throughput (and associated labour costs), low genotyping error rates and ease of data standardization across experiments and laboratories. In our laboratory, a single researcher can obtain data for ~41 000 single-locus genotypes per week with fully automated scoring and ~44 000 genotypes per week with some manual editing. We estimate this to be at least 40-fold more efficient than traditional methods. As noted, genotyping error rates are low and comparable to those obtained with carefully selected loci using conventional electrophoretic methods in other studies (Hoffman & Amos 2005; Pompanon *et al.* 2005; Hess *et al.* 2012). As genotypes are based on direct counts of DNA sequence lengths rather than indirect inference from electrophoretic data, data standardization between platforms and laboratories is not a concern.

Third, notwithstanding the benefits of fully automated genotype scoring, there will be situations where manual editing is desirable. For example, it may be advantageous to include somewhat more difficult to score loci to enable comparisons with older data sets, or comparisons across species, or because particular loci have particular merits, such as being linked to genes or traits of interest. The data visualization feature in MEGASAT enables easy manual checking and editing of genotypes, and our results suggest that rapid manual editing can improve genotyping accuracy at loci that might otherwise be of marginal utility. Conversely, although the default values of variables MEGASAT uses to guide the decision making process for identifying true alleles among amplification or sequencing artefacts work well for a wide variety of di- and trinucleotide microsatellite loci, locus-specific adjustments of some of these user-definable variables may improve the automated scoring accuracy of some problematic loci.

In addition to the benefits of high throughput and high reproducibility of genotype calling, sequencing-based genotyping of microsatellites entails at least four further advantages: first, genotyping costs can be very low. Costs will vary depending on factors such as the number of loci in multiplexes, but for example, in our study of 43 microsatellite loci, consumable and sequencing costs are less than $0.04 (U.S.) per locus, which compares favourably with the cost of many SNP assays, particularly when the greater per locus information content of microsatellites is considered. Second, the fact that the compositions of PCR multiplexes are not constrained by concerns about size overlaps among loci (as would be the case for electrophoresis-based genotyping) allows for many loci to be combined in single multiplexes. This not only contributes to savings in genotyping costs, but also reduces the amount of DNA needed as well. We routinely genotype our 43 loci using subnanogram quantities of DNA. Third, less optimization of PCR multiplexes is required for sequencing-based genotyping than for electrophoresis-based methods. This is primarily because nontarget amplicons (which can complicate interpretation of electrophoretic data) are automatically filtered out by MEGASAT, but also because the extremely large dynamic range of sequencing-based detection makes wide variation in amplification efficiency acceptable. The NGS-based approach to microsatellite genotyping brings considerable flexibility to the genotyping process. Loci can be added or removed from the marker panel with relatively little effort or complications.

We conclude with two observations of general relevance to the utility of microsatellites and MEGASAT. The first is that the advent of NGS has caused the cost of microsatellite discovery to plummet. A few hundred dollars' worth of NGS data for any given species will usually lead to discovery of hundreds-to-thousands of microsatellite sequences that hold potential as genetic markers (e.g. Gardner *et al.* 2011; Guichoux *et al.* 2011; Souza *et al.* 2015). Importantly, as microsatellite polymorphism is closely correlated with the number of repeats (up to about 10 repeats), the likelihood that a

given microsatellite sequence is polymorphic can be predicted directly from sequence data obtained from a single individual (e.g. Payseur *et al*. 2011). This reduces both the likelihood of ascertainment bias relative to SNPs and the overall cost of marker development. Flexible multiplex PCRs, avoidance of gel electrophoresis and wide latitude in acceptable PCR results (because of sequence based data filtering) make it relatively easy to convert microsatellite sequence data to working microsatellite markers. For example, in a recent study (in prep.), we achieved >60% conversion of microsatellite sequences (from genomic data) to working microsatellite loci, with no optimization of PCR parameters (117 of 192 loci, data not shown). Finally, studies which require only dozens-to-hundreds of loci (whether microsatellites or SNPs), and especially studies where DNA quantity or quality are limited, will benefit from analysis of microsatellites using MEGASAT with NGS data for highly cost-effective acquisition of genetic data.

Acknowledgements

This research benefitted from a Canadian Natural Sciences and Engineering Research Council (NSERC) Strategic Grant to PB and RGB, an NSERC Discovery Grant to PB and National Science Foundation (NSF) support to DR. The sequence data reported in this study were obtained using a DNA sequencer acquired using a generous bequest from Elizabeth Ann Nielsen.

References

Aime C, Verdu P, Segurel L *et al*. (2014) Microsatellite data show recent demographic expansions in sedentary but not in nomadic human populations in Africa and Eurasia. *European Journal of Human Genetics*, **22**, 120101207.

Campbell NR, Harmon SA, Narum SR (2014) Genotyping-in-Thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, **15**, 855–867.

Faircloth BC (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources*, **8**, 92–94.

Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011) Rise of the machines – recommendations for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources*, **11**, 1093–1101.

Guichoux E, Lagache L, Wagner S *et al*. (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources*, **11**, 591–611.

Haasl RJ, Payseur BA (2010) The number of alleles at a microsatellite defines the allele frequency spectrum and facilitates fast, accurate estimation of theta. *Molecular Biology and Evolution*, **27**, 2702–2715.

Hess MA, Rhydderch JG, LeClair LL *et al*. (2012) Estimation of genotyping error rate from repeat genotyping, unintentional recaptures and known parent-offspring comparisons in 16 microsatellite loci for brown rockfish (*Sebastes auriculatus*). *Molecular Ecology Resources*, **12**, 1114–1123.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, **14**, 599–612.

Jarne P, Lagoda P (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution*, **11**, 424–429.

Jones OR, Wang J (2010) COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, **10**, 551–555.

Kearse M, Moir R, Wilson A *et al*. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Kimpton CP, Gill P, Walton A *et al*. (1993) Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Methods and Applications*, **3**, 13–22.

Li JZ, Absher DM, Tang H *et al*. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–1104.

Moran P, Teel DJ, LaHood ES *et al*. (2006) Standardising multi-laboratory microsatellite data in Pacific salmon: an historical view of the future. *Ecology of Freshwater Fish*, **15**, 597–605.

Payseur BA, Jing P, Haasl RJ (2011) A genomic portrait of human microsatellite variation. *Molecular Biology and Evolution*, **28**, 303–312.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, **6**, 847–859.

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology and Evolution*, **16**, 142–147.

Putman AI, Carbone I (2014) Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecology and Evolution*, **4**, 4399–4428.

Seeb JE, Pascal CE, Ramakrishnan R, Seeb LW (2009) SNP genotyping by the 5′-nuclease reaction: advances in high throughput genotyping with non-model organisms. In: *Methods in Molecular Biology, Single Nucleotide Polymorphisms* (ed. Komar A), pp. 277–292. Humana Press, New York.

Souza IGB, Paterson I, McBride MC *et al*. (2015) Isolation and characterization of 23 microsatellite loci in the stingless bee *Melipona subnitida* using next generation sequencing. *Conservation Genetics Resources*, **7**, 239–241.

Suez M, Behdenna A, Brouillet S *et al*. (2016) MicNeSs: genotyping microsatellite loci from a collection of (NGS) reads. *Molecular Ecology Resources*, **16**, 524–533.

Sun JX, Mullikin JC, Patterson N, Reich DE (2009) Microsatellites are molecular clocks that support accurate inferences about history. *Molecular Biology and Evolution*, **26**, 1017–1027.

Wright JM, Bentzen P (1994) Microsatellites: genetic markers for the future. *Reviews in Fish Biology and Fisheries*, **4**, 384–388.

L.Z. created the MEGASAT software. I.P. implemented the laboratory protocols, performed the laboratory work and assigned the microsatellite genotypes. B.F. provided sequence data for many hundreds of unpublished guppy microsatellites, including the 43 that ended up being used in this project. B.W. carried out the analyses of genotyping error rates. I.B. contributed to project supervision and editing the manuscript. P.N.R. assisted with development of the MEGASAT software. D.R., provided the guppy samples and is PI for the project that led to the need for MEGASAT. R.B. supervised and contributed to the development of the MEGASAT software and contributed to writing the manuscript. P.B. led the MEGASAT project and wrote much of the manuscript.

Data accessibility

The MEGASAT software, user manual and sample data set are freely available at https://github.com/beiko-lab/MEGASAT.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Accession numbers and sequences of 448 *Poecilia reticulata* microsatellite loci.

Appendix S2 Summary of the lab work and sequencing runs used to create the DNA sequence input for MEGASAT.

Appendix S3 Example of relative position of alleles. Megasat uses different decision rules based on relative position of the dominant alleles.

Appendix S4 Error Rates per locus for pedigree and re-genotyping data.